

Missing values and lack of information in water management datasets: an approach based on Bayesian Networks

Rosa F. Roperó¹, M. Julia Flores², Rafael Rumí¹

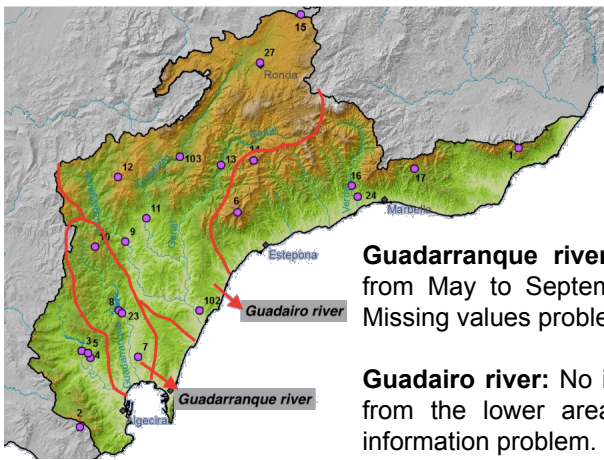
¹Data Analysis Research group, Dpt. Mathematics, University of Almería, Spain

²Computing Systems Department, SIMD I3A, University of Castilla-la Mancha, Campus Univ., Albacete, Spain

Introduction

Environmental data often present missing values or lack of information that make modelling tasks difficult. Under the framework of SAICMA Research Project, a flood risk management system is modelled for Andalusian Mediterranean catchment using information from the Andalusian Hydrological System. Hourly data were collected from October 2011 to September 2020.

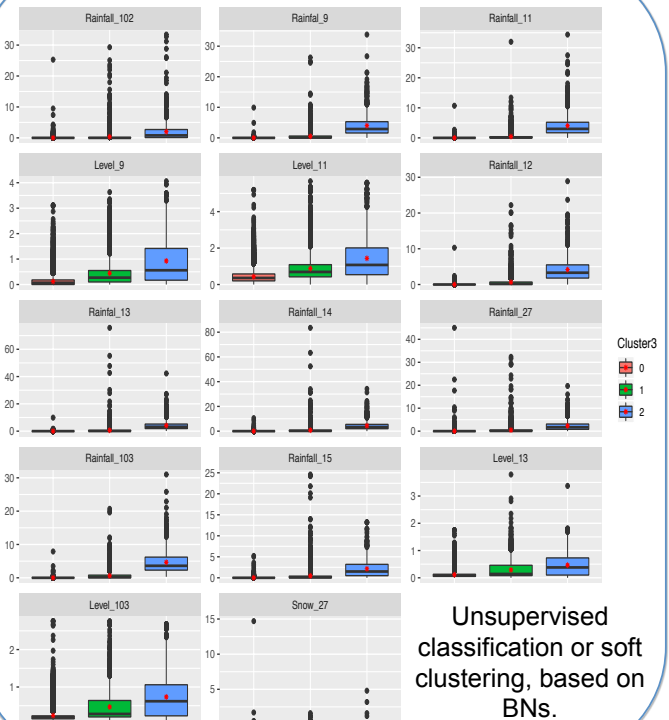
Study area



Guadarranque river: No data from May to September 2020. Missing values problem.

Guadaira river: No information from the lower area. Lack of information problem.

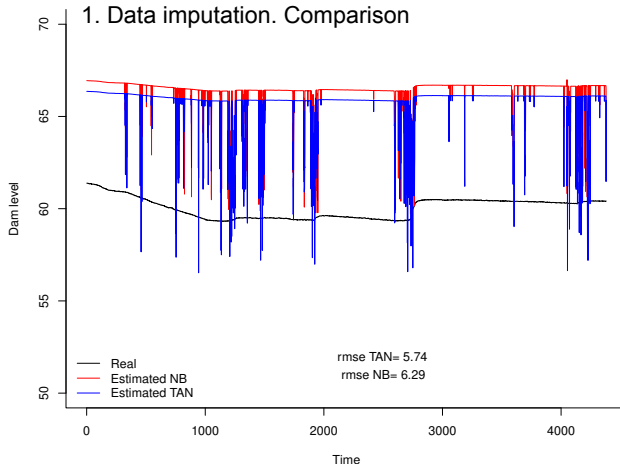
Guadaira river



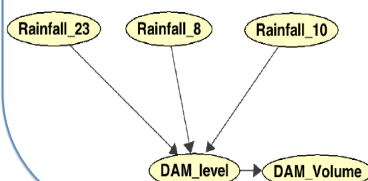
Unsupervised classification or soft clustering, based on BNs.

Guadarranque river

1. Data imputation. Comparison



2. Model learned with real and imputed data.



Model	rmse
Real	4.26
NB imputed	3.33
TAN imputed	3.07

Conclusion and Future Works

Solutions based on Bayesian networks provide appropriate and robust results. But, further work is needed:

- Scenario with new data to check if extreme events are properly identified.
- Include information about consumption rates in order to improve final predictions.

Acknowledgements

This study was supported by:

- Regional Government of Andalusia, project SAICMA (UAL18-TIC-A011- B-E)
- Spanish Ministry of Economy and Competitiveness through projects: PID2019-106758GB-C32, PID2019-106758GB-C33, TIN2016-77902-C3-1-P and TIN2016-77902-C3-3-P.